GridICE: architecture and sensor at fabric level

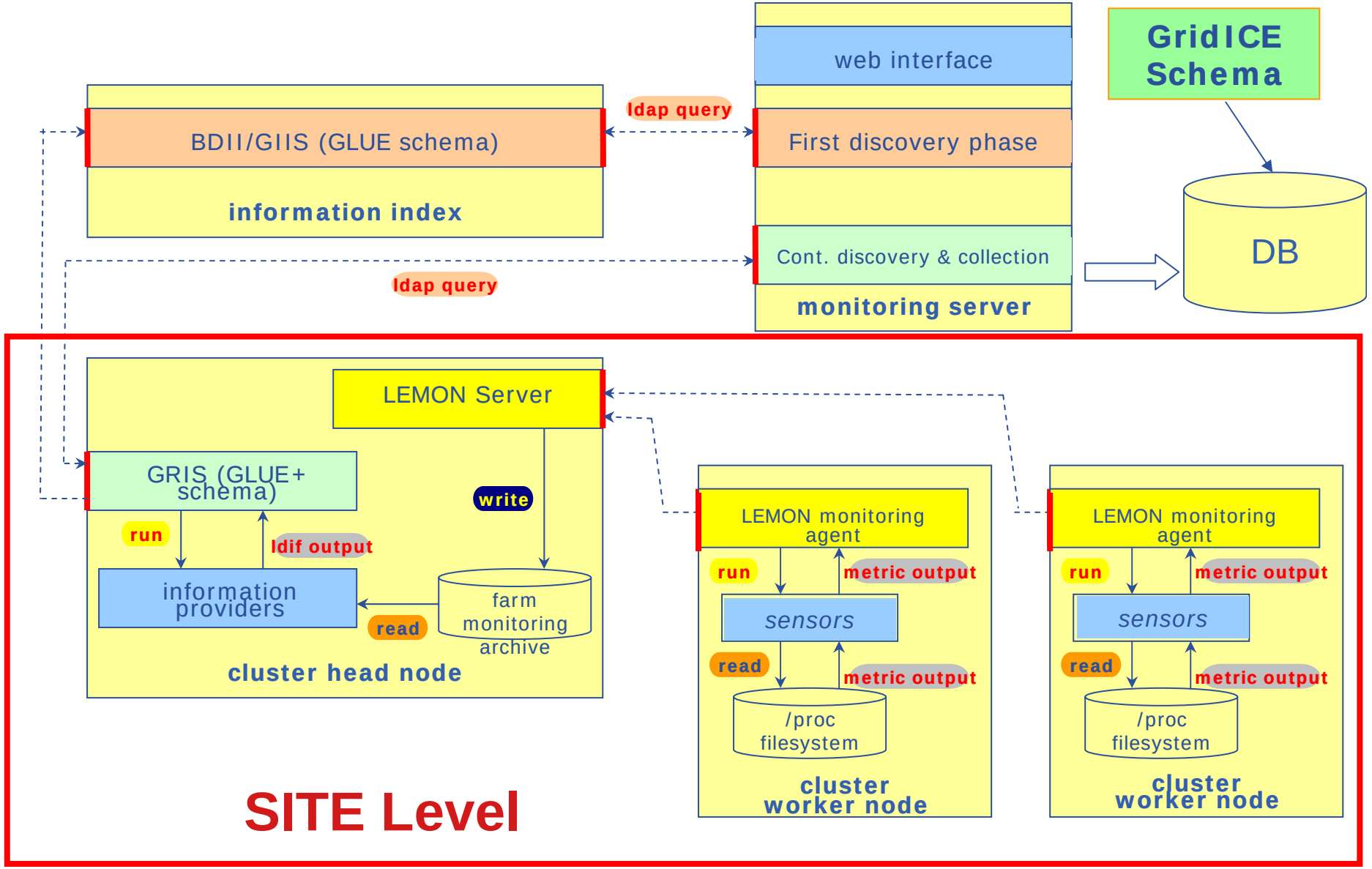
Sergio Fantinel
INFN (Italy)
sergio.fantinel@lnl.infn.it

Grid Monitoring WG @ WLCG Collaboration Workshop
2007-01-25

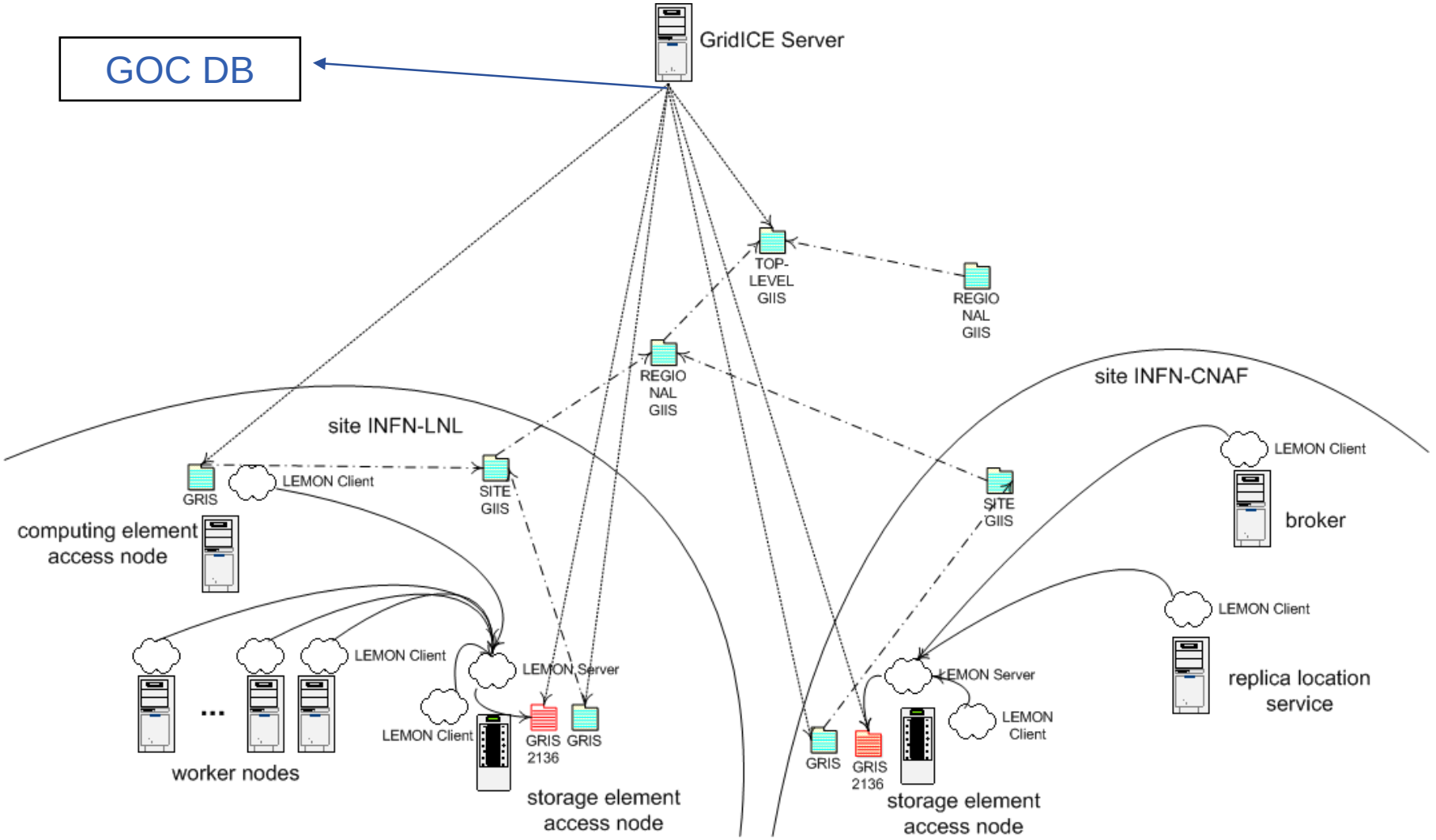
OUTLINE

- Brief Architecture Overview
- Fabric Level Sensors
- Introducing to the GUI and related data collected on the concentrator (GridICE Server)

Infrastructure Layout



Deployment Layout and info Sources



GridICE sensors

- **Standard per host metric collection**
  - CPU, MEM, VMEM, FileSystem, Net Interfaces, …
- **GridICE specific probes for advanced measurement**
  - In Production:
    - Job monitoring
    - LRMSInfo
    - Generic/per role Daemons status with simple config
  - Ready but not in release:
    - WMSLB
  - In developing (almost ready)
    - storage (CASTOR, DPM, dCache) internal status (almost done, integration needed)
    - FTS

Daemons Sensor (I)

- **Returns several metrics describing the status of a given list of daemons**

Lemon Name: CheckDaemon			**Lemon ID**: 00010104
Name	**Unit**	**DataType**	**Description**
role	-	string[255]	Role name
daemon_name	-	string	Daemon name to check
commandline	-	string[1024]	String which reports the process checked
cpuusageall	%	int	Total cpu percentage for all the istances of the process
cpuusageonemax	%	int	Max cpu percentage among the istances that process
firststarted	sec	string[50]	First process istance elapsed time
laststarted	sec	string[50]	Last process istance elapsed time
memusageavg	%	int	Average percentage memory usage by the istances of the process
memusageonemax	%	int	Max percentage memory usage among the istances of the process
numofinstances	-	int	Numeber of istances of a process
status	-	string[20]	Status of the process
timeusageall	sec	string[50]	Accumulated cpu time (user + system) by all istances
timeusageonemax	sec	string[50]	Max accumulated cpu time (user + system) among all the instances

Daemons Sensor (II)

- The list of the daemons to be monitored described in a simple configuration text-file
- Can be changed runtime at any moment!!

```
[root@t2-ce-02 root]# cat /opt/gridice/monitoring/etc/gridice-role.cfg
[ce-access-node]
gsiftp ^[\s\w\/\.-]*ftpd
edg-gatekeeper ^[\s\w\/\.-]*edg-gatekeeper
globus-mds ^[\s\w\/\.-]*/opt/globus/libexec/slapd
fmon-agent ^[\s\w\/\.-]*fmon-agent
lcg-bdii-fwd     ^[\s\w\/\.-]*bdii-fwd
lcg-bdii-update ^[\w\/\.-]*perl\s[\s\w\/\.-]*bdii-update
lcg-bdii-slapd  ^[\w\/\.-]*slapd\s[\s\w\/\.\-]*bdii
dgas-pushd ^[\s\w\/\.-]*glite-dgas-pushd
dgas-gianduia ^[\s\w\/\.-]*glite-dgas-gianduia
dgas-ceserverd ^[\s\w\/\.-]*glite-dgas-ceserverd
dgas-had ^[\s\w\/\.-]*glite-dgas-ceServerd-had
gridice_messlog ^[\s\w\/\.-]*messlog_mon
gridice_lsf ^[\s\w\/\.-]*parse_lsf
lsf-lim ^[\s\w\/\.-]*lim
lsf-pim ^[\s\w\/\.-]*pim
lsf-res ^[\s\w\/\.-]*res
lsf-sbatchd ^[\s\w\/\.-]*sbatchd
[ce-access-node end]
[root@t2-ce-02 root]#
```

Daemons Sensor (III)

- **What you see at server side**

INFN-LNL-2 >> Host::t2-ce-02.lnl.infn.it

XML

Role	Proc Name	Status	Inst#	First	Last	CPU1Max	CPUAll	Mem1Max	MemAvg	Time1Max	TimeAll
ce-access-node	dgas-ceserverd	STOP	0	0-00:00	0-00:00	0	0	0	0	0-00:00	0-00:00
ce-access-node	dgas-gianduia	STOP	0	0-00:00	0-00:00	0	0	0	0	0-00:00	0-00:00
ce-access-node	dgas-had	S	1	7-20:29	7-20:29	0	0	0	0	0-00:01	0-00:01
ce-access-node	dgas-pushd	S	1	7-20:52	7-20:52	1	1	0	0	0-01:48	0-01:48
ce-access-node	edg-gatekeeper	S	1	13-23:01	13-23:01	0	0	0	0	0-00:00	0-00:00
ce-access-node	fmon-agent	S	1	36-19:21	36-19:21	0	0	0	0	0-00:20	0-00:20
ce-access-node	globus-mds	S	2	0-00:00	0-00:00	0	0	0	0	0-02:20	0-02:20
ce-access-node	gridice_lsf	S	2	36-19:25	36-19:25	0	0	0	0	0-00:14	0-00:26
ce-access-node	gridice_messlog	S	1	36-21:25	36-21:25	0	0	3	3	0-00:21	0-00:21
ce-access-node	gsiftp	S	1	13-21:41	13-21:41	0	0	0	0	0-00:00	0-00:00
ce-access-node	lcg-bdii-fwd	S	3	33-03:33	0-00:00	0	0	0	0	0-00:52	0-00:52
ce-access-node	lcg-bdii-slapd	S	3	0-00:01	0-00:00	1	1	0	0	0-00:00	0-00:00
ce-access-node	lcg-bdii-update	S	1	33-01:34	33-01:34	0	0	0	0	0-00:24	0-00:24
ce-access-node	lsf-lim	S	1	0-00:00	0-00:00	0	0	0	0	0-04:20	0-04:20
ce-access-node	lsf-pim	S	1	0-00:00	0-00:00	0	0	0	0	0-00:01	0-00:01
ce-access-node	lsf-res	S	2	0-00:00	0-00:00	0	0	0	0	0-00:02	0-00:02
ce-access-node	lsf-sbatchd	S	1	0-00:00	0-00:00	0	0	0	0	0-00:01	0-00:01

Site can be notified by the GridICE server on daemons status change

Daemons Sensor (IV)

- **With the MW we distribute a number of predefined roles (each different role has its own list of monitored daemons):**
  - Lcg-CE-LSF*
  - Lcg-CE-PBS*
  - gLite-CE-LSF*
  - gLite-CE-PBS*
  - SE-CLASSIC
  - DPM-HeadNode
  - DPM-PoolDisk
  - dCache-HeadNode
  - dCache-PoolDisk
  - MONBOX
  - BDII
  - lcg-RB
  - gLite-WMSLB
  - HLR**
  - WN-LSF**
  - WN-PBS**

* Different list for gLite/INFNGRID (DGAS)
** Only on INFNGRID

- **A site Administrator can modify the defaults and can invent new roles (PhEDEX, UI, LRMS-SERVER, per VO-BOX, FTS,…), they will be tracked by the server**

Job Monitoring Sensor

- **Use daemons paradigm to collect information from different log files and other sources (LRMS) -> feed a cache**
- **A simple probe gather the cached info and push them into LeMON**

LocalID	-	int	Local batch system job id
Type	-	string[255]	Batch system name (pbs,lsf...)
HostUniqueID	-	string[255]	Computing Element hostname
GlobalID	-	string[255]	Grid job id
Status	-	char	Job status (Q,R,E)
Name	-	string[255]	Job name
LocalOwner	-	string[32]	Local account user is mapped to
GlobalOwner	-	string[255]	User certificate DN
ExecutionTarget	-	string[255]	Execution host (for jobs R and E)
CPUTime	s	long int	Cputime used by job (for jobs R and E)
WallTime	s	long int	Walltime used by job (for jobs R and E)
ExitStatus	-	int	Signed code for jobs exit status
RAMUsed	KByte	long int	RAM used by job
VirtualUsed	KByte	long int	Virtual memory used by job
CreationTime	s	long int	Queued time of job (since Unix age)
StartTime	s	long int	Start time of job (for R,E jobs - since Unix epoch)
EndTime	s	long int	End time of job (for E jobs - since Unix epoch)
JobQueue	-	string	Queue name where's queued job
JobVO	-	String	User vo name

LRMSInfo Sensor

- **Try to avoid double counting of resources inspecting the configuration and status of the LRMS**

Name	Unit	DataType	Description
HostUniqueID	-	string[255]	Computing Element hostname
Type	-	string[255]	Batch system name (pbs,lsf...)
TotalJobSlots	job slot	int	Total number of logical cpus
FreeJobSlots	job slot	int	Total number of free logical cpu
WaitingJobs	job	int	Number of queued jobs
NodeCount	node	int	Number of worker nodes
CPULoadAvg	process	int	Average of cpu loads of all nodes (multiplied by 100)
RAMTotal	MByte	long int	Sum of total RAM (swap included) of all WNs
RAMUsed	MByte	long int	Sum of total used RAM of all WNs
NodeDownCount	node	int	Number of nodes not available (down,offline...)

WMS Sensor

- **Ready for gLite WMS, not in release yet. Easy to modify to work with lcg-RB (not LeMON output format yet, simple to do)**
- **Ref. URL: http://goc.grid.sinica.edu.tw/gocwiki/RB_Passive_Sensor**

WMS_Sensor_Version
CG_EndedJobs1H
CG_HeldJobs
CG_RunningJobs
CG_SubmittedJobs1H
CG_WaitingJobs
JC_InputFileListSize
JC_WaitingRequests
WMS_Jobs_Aborted
WMS_Jobs_Cancelled
WMS_Jobs_Cleared
WMS_Jobs_Done
WMS_Jobs_Purged
WMS_Jobs_Ready
WMS_Jobs_Running
WMS_Jobs_Scheduled
WMS_Jobs_Submitted
WMS_Jobs_Unknown
WMS_Jobs_Waiting
WMS_Jobs_Aborted1H
WMS_Jobs_Cancelled1H
WMS_Jobs_Cleared1H
WMS_Jobs_Done1H
WMS_Jobs_Purged1H
WMS_Jobs_Ready1H
WMS_Jobs_Running1H
WMS_Jobs_Scheduled1H
WMS_Jobs_Submitted1H
WMS_Jobs_Unknown1H
WMS_Jobs_Waiting1H
WMS_SandBox_InputSandBoxMaxSize
WMS_SandBox_InputSandBoxNumber
WMS_SandBox_InputSandBoxSizeTotal
WMS_SandBox_OutputSandBoxMaxSize
WMS_SandBox_OutputSandBoxNumber
WMS_SandBox_OutputSandBoxSizeTotal
WM_InputFileListSize
WM_WaitingRequests

- **Components:**
  - **WM_* = WorkLoad Manager**
    WMS_* = Whole System
    CG_* = Condor-G
    JC_* = Job Controller

Advanced Storage Sensors (I)

CASTOR (gsiftp, rfio), **DPM** (gsiftp, rfio), **dCache** (gsiftp, dcap)

- **Operation type**
  - *Read* o *Write*
  - *Access protocol*
  - *Local/remote access*
- **Transferred files**
  - *Filename* (Full path)
  - *Byte transferred*
  - *Streams number*
  - *Exit_status*
- **Used hosts**
  - *Source machine*
  - *Dest machine*
  - *Submit machine*
- **Timings**
  - *Start* (local time)
  - *End* (local time)
  - *Duration*
  - *Shift* (UTC)
- **Detailed user info**
  - *Local user*
  - *VO*
  - *DN* (write operation)
  - *DN* (read operation)

Advanced Storage Sensors (I)

CASTOR (gsiftp, rfio), **DPM** (gsiftp, rfio), **dCache** (gsiftp, dcap)

- **Operation type**
  - *Read* o *Write*
  - *Access protocol*
  - *Local/remote access*
- **Transferred files**
  - *Filename* (Full path)
  - *Byte transferred*
  - *Streams number*
  - *Exit_status*
- **Used hosts**
  - *Source machine*
  - *Dest machine*
  - *Submit machine*
- **Timings**
  - *Start* (local time)
  - *End* (local time)
  - *Duration*
  - *Shift* (UTC)
- **Detailed user info**
  - *Local user*
  - *VO*
  - *DN* (write operation)
  - *DN* (read operation)

Advanced Storage Sensors (II)

- **The Advanced Storage Sensors for monitoring and accounting are almost done:**
  - **CASTOR**: waiting CERN for a fix in rfio log (configuration?) reporting. Tested at INFN-T1
  - **DPM**: code almost ready, only few changes, need. testing
  - **dCache**: in test since some months at INFN-BARI. Overall good results

- **(not LeMON output format yet)**

FTS Sensor

- **In test at INFN-T1 FTS server**
- **The developing is stopped because of the imminent update of the service**

Measurement Class: FTS_jobs

job_id	-	string[150]	Job id
jobstatus	-	string[10]	Regarding if the job is active or finished.
job_user	-	text	The user credentials
source_srm	-	text	The link of the source srm
dst_srm	-	text	The link of the destination srm
source	-	text	The actual source link
destination	-	text	The actual destination link
start_time	-	int[11]	Time when the job was submitted.
end_time	-	int[11]	Time when the jobs finished.(In the case of active jobs This will be the time when the query was done)
trans_time	-	int[11]	Time taken for the transfer of the file.
rate	-	double	The rate of the transfer (bites/sec).
jobbytes	-	int[11]	The size of the file which was transferred

All LeMON compatible Sensors

- **Measurement specification for GridICE-LeMON compatible probes are provided for:**
  - CPUINFO (10100)
  - OS (10101)
  - ALIVE (10102)
  - REGFILES (10103)
  - DAEMON (10104)
  - UPTIME (11001)
  - CPU (11011)
  - MEMORY (11021)
  - SWAP (11022)
  - PROCESSES (11031)
  - DISK (11101)
  - SOCKETS (11201)
  - NETWORK (11202)
  - Jobs (10106)
  - LRMSInfo (10107)